Blind Signal Separation in the Presence of Gaussian Noise

Mikhail Belkin
Ohio State University
Computer Science and Engineering,
2015 Neil Avenue, Dreese Labs 597.
Columbus, OH 43210
mbelkin@cse.ohio-state.edu

Luis Rademacher
Ohio State University
Computer Science and Engineering,
2015 Neil Avenue, Dreese Labs 495.
Columbus, OH 43210
lrademac@cse.ohio-state.edu

James Voss
Ohio State University
Computer Science and Engineering,
2015 Neil Avenue, Dreese Labs 586.
Columbus, OH 43210
vossj@cse.ohio-state.edu

November 9, 2012

Abstract

A prototypical blind signal separation problem is the so-called *cocktail party problem*, with n people talking simultaneously and n different microphones within a room. The goal is to recover each speech signal from the microphone inputs. Mathematically this can be modeled by assuming that we are given samples from a n-dimensional random variable $\mathbf{X} = A\mathbf{S}$, where $\mathbf{S}$ is a vector whose coordinates are independent random variables corresponding to each speaker. The objective is to recover the matrix A^{-1} given random samples from $\mathbf{X}$. A range of techniques collectively known as Independent Component Analysis (ICA) have been proposed to address this problem in the signal processing and machine learning literature. Many of these techniques are based on using the kurtosis or other cumulants to recover the components.

In this paper we propose a new algorithm for solving the blind signal separation problem in the presence of additive Gaussian noise, when we are given samples from $\mathbf{X} = A\mathbf{S} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is drawn from an unknown n-dimensional Gaussian distribution. Our approach is based on a method for decorrelating a sample with additive Gaussian noise under the assumption that the underlying distribution is a linear transformation of a distribution with independent components. Our decorrelation routine is based on the properties of cumulant tensors and can be combined with any standard cumulant-based method for ICA to get an algorithm that is provably robust in the presence of Gaussian noise. We derive polynomial bounds for sample complexity and error propagation of our method.

Our results generalize the recent work of Arora et al. [1] which deals with a special case of ICA when $\mathbf{S}$ is the uniform probability distribution over the binary cube.

1 Introduction and related work

A prototypical blind signal separation setting is the so-called *cocktail party problem*: in a room, there are n people speaking simultaneously and n microphones, with each microphone capturing a superposition of the voices. The objective is to recover the voice of each individual speaker. The

simplest modeling assumption is to consider each speaker as producing a signal to be a random variable independent of the others and to take the superposition to be a linear transformation independent of time. This leads to the following problem: given a sample from n-dimensional random variable $\mathbf{X}$, satisfying $\mathbf{X} = A\mathbf{S}$, where A is a non-singular square matrix and $\mathbf{S}$ is another random vector whose coordinates are unknown independently distributed (but not necessarily identical) random variables, we need to recover the matrix A^{-1}. Equivalently, we need to recover the basis corresponding to the directions of the independent components.

The name Independent Component Analysis refers to a broad range of algorithms addressing this signal separation problem as well as its variants and extensions. It has generated significant interest and an extensive literature in the signal processing and machine learning communities due to its applicability to a variety of important practical situations including speech [13], vision [2] and various biological and medical applications, e.g., [11]. For a comprehensive introduction see the books [4, 10].

One widely used class of algorithms for ICA is based on the remarkable fact that if the data is whitened, that is, $\mathbf{X}$ has the zero mean and the identity covariance matrix, then the absolute value of kurtosis reaches its maximum in the directions corresponding to the independent components. More precisely, consider the kurtosis as a function on the n-dimensional unit sphere. For whitened data it can be defined as follows:

$$\mathbf{v} \mapsto \kappa_4(\mathbf{v} \cdot \mathbf{X}) := \mathbb{E}((\mathbf{v} \cdot \mathbf{X})^4) - 3$$

It can be shown [5, 6] that the vectors corresponding to the maxima of the absolute value of $\kappa_4(\mathbf{v} \cdot \mathbf{X})$ form an orthonormal basis whose elements are independent random variables. Thus the underlying structure of the signal can be recovered by analyzing the behavior of this function. Moreover, computing the kurtosis involves the expected value of the fourth power of a random variable, which can be easily approximated from a finite sample.

This observation leads to the following procedure for the Independent Component Analysis in the noiseless case:

Step 1. "Whiten" the original signal, that is apply a linear transformation which transforms the covariance matrix of the sample to the identity. This is typically achieved by using the Principal Component Analysis (PCA) to transform the input data to the basis of its principal directions by an orthogonal transformation and rescaling the resulting data appropriately.

Step 2. After the signal is whitened various optimization procedures can be used to find the maxima of the absolute value of kurtosis over the unit sphere. The independent components are recovered from the directions of these maxima.

In their recent paper [1] Arora, et al. make an important observation that for a slight variation of Step 2 to work, it is sufficient for the the sample to be decorrelated (*quasi-whitened*), that is, to have independent coordinates in some orthogonal basis, rather then fully whitened (having the identity covariance matrix).

In this paper we consider the problem of signal separation for a noisy signal $\mathbf{X} = A\mathbf{S} + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an unknown n-dimensional Gaussian distribution. The main difficulty is in Step 1, since the principal directions given by PCA are contaminated by the noise and do not generally decorrelate the underlying signal. Interestingly, as a result of the invariance of the kurtosis under the additive Gaussian noise, Step 2 of the algorithm is still valid and the usual methods and analyses still apply with minor caveats.

The main contribution of our paper is addressing the problem of decorrelating the underlying signal in the presence of noise. We show how to approximate a matrix B, such that $B^{-1}A$ is diagonal in the basis of independent coordinates. We provide polynomial bounds for the sample complexity and error analysis as well as an analysis of error propagation compatible with any analysis of Step 2.

Our approach can be viewed as a noise-invariant version of PCA for the special case when the underlying probability distribution is a product of independent variables. The method is based on

the properties of the fourth cumulant tensor, rather than the usual covariance matrix used in PCA. To the best of our knowledge, this is the first general algorithm for noisy ICA with sample complexity and running time guarantees. Moreover, unlike methods such as [20], our approach is compatible with any optimization procedure for the Step 2.

Related work. Over the last twenty years blind signal separation[1] has become a large and active area of research in signal processing and machine learning community. An important class of methods for ICA is based on the properties of kurtosis and other higher-order cumulants.

Most of these works concentrate on algorithms, implementations and applications and do not provide a sample or running-time complexity analysis for the algorithms. One such analysis is provided in Frieze, et al. [6], where the authors address the question of learning a linear transformation, which is equivalent to the ICA problem, and provide a complexity analysis. In a slightly different context of cryptoanalysis, [16] analyzes a kurtosis-based method method for learning a parallelopiped. In [19] the authors analyze a generalized version of ICA for learning higher-dimensional subspace "juntas" in the presence of noise.

The problem of blind signal separation in the presence of noise has been an active topic of research in the machine learning literature. In particular we would like to point out the work of Yeredor [20] which proposes an elegant one-step approach for general ICA with Gaussian noise, based on approximating the Hessian of the second characteristic function, namely $v \mapsto \nabla_v^2 \log \mathbb{E}_x(e^{v^T x})$, at a finite number of generic choices of v. The recent work of Hsu and Kakade [7, Section 3, Theorem 3] proposes an approach similar to Yeredor's, using the Hessian of the directional kurtosis instead of the second characteristic function and makes interesting connections to learning Gaussian mixture distributions in high dimension. Finally, the sample complexity analysis for noisy ICA is studied in Arora et al. [1], which provides a complete discussion for the special case when the underlying signal is a uniform distribution over the n-dimensional binary cube $\{-1, 1\}^n$.

We also would like to point out that our approach is closely related to the class of tensor methods for data analysis, see e.g. [17, 15]

2 Properties of Cumulants

Let $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^T\mathbf{X})]$, $\mathbf{t} \in \mathbb{R}^n$ denote the first charateristic function of a n-dimensional vector valued random variable $\mathbf{X}$, and let $\psi_{\mathbf{X}}(\mathbf{t}) = \log(\phi_{\mathbf{X}}(\mathbf{t}))$ denote the second characteristic function of $\mathbf{X}$. Cumulants are defined as the coefficients of the Taylor Expansion of the second characteristic function. Specifically, using the multi-index notation, we write

$$1 + \sum_{r=1}^{\infty} \sum_{i_1,\cdots,i_n \in [n]^r} \frac{1}{r!} i^r \left(\prod_{j=1}^{r} \mathbf{t}_{i_j} \right) \mathrm{Cum}(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_r}) = \psi_{\mathbf{X}}(\mathbf{t}).$$

For each cumulant $\mathrm{Cum}(\mathbf{X}_{i_1}, \cdots, \mathbf{X}_{i_r})$, r is referred to as the order of the cumulant. Order r cumulants of a random variable $\mathbf{X}$ can be collected into a cumulant tensor, called the r^{th} cumulant tensor of $\mathbf{X}$. For instance, the fourth order cumulant tensor of $\mathbf{X}$, denoted by $Q_{\mathbf{X}}$ in this paper, is defined by $(Q_{\mathbf{X}})_{ijkl} = \mathrm{Cum}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l)$. Since any simultaneous draw of random variables can be viewed as a draw of a single vector-valued random variable, this definition can be used to construct cross cumulants between arbitrary random variables. In the univariate case in which X and t are scalars, the notation $\kappa_r(X)$ is used to denote the r^{th} order cumulant $\mathrm{Cum}(X, \cdots, X)$.

Cumulants are similar in flavor to moments, and indeed all cumulants have polynomial expansions in terms of the moments of the same and smalle order. For example, the fourth cumulant (kurtosis) of a 0-mean one-dimensional random variable X can be expanded $\kappa_4(X) = \mathbb{E}[X^4] - 3\mathbb{E}[X^2]^2$. However, cumulants have nice algebraic properties not shared by moments, properties on which this work

[1] Also known as Blind Source Separation.

relies heavily. Let $X_1, \cdots, X_r$ be real-valued random variables. Then, cross cumulants are known to manifest the following properties:

1. (Multilinearity) If $c_i \in \mathbb{R}$ is a constant, then

$$\text{Cum}(X_1, \cdots, c_i X_i, \cdots, X_r) = c_i \text{Cum}(X_1, \cdots, X_i, \cdots, X_r).$$

   Also, if Y_i is a random variable, then

$$\begin{aligned}&\text{Cum}(X_1, \cdots, X_i + Y_i, \cdots, X_r)\\ &\quad = \text{Cum}(X_1, \cdots, X_i, \cdots, X_r) + \text{Cum}(X_1, \cdots, Y_i, \cdots, X_r).\end{aligned}$$

2. (Independence) If 2 variables X_i and X_j $(i < j)$ are independent random variables, then the cross cumulant $\text{Cum}(X_1, \cdots, X_i, \cdots, X_j, \cdots, X_n)$ is zero. Combined with the multilinearity property, this implies that if the variables $Y_1, \cdots, Y_n$ are independent of $X_1, \cdots, X_n$, then

$$\begin{aligned}&\text{Cum}(X_1 + Y_1, X_2 + Y_2, \cdots, X_n + Y_n)\\ &\quad = \text{Cum}(X_1, X_2, \cdots, X_n) + \text{Cum}(Y_1, Y_2, \cdots, Y_n).\end{aligned}$$

3. (Vanishing Gaussians) The only non-zero cumulant tensors of Gaussian random variables are the 1-tensor mean and the 2-tensor covariance matrix.

Note that in the univariate case, these properties become:

1. (Additivity) If X and Y are independent random variables, then $\kappa_r(X+Y) = \kappa_r(X) + \kappa_r(Y)$.
2. (Homogeneity of degree r) If c is a constant, then $\kappa_r(cX) = c^r \kappa_r(X)$.
3. (Vanishing Gaussians) The only non-zero cumulants of a Gaussian random variable are the mean and the variance (the first and second order cumulants).

3 Problem Statement and Main Result

Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)} \in \mathbb{R}^n$ be an i.i.d. N-sample of vector-valued random variables. In independent component analysis (ICA) it is assumed that each $\mathbf{x}^{(i)}$ is generated from a latent random variable $\mathbf{s}^{(i)}$ via an unknown mixing matrix A such that

$$\mathbf{x}^{(i)} = A\mathbf{s}^{(i)} + \boldsymbol{\eta}^{(i)}$$

where $\boldsymbol{\eta}$ is additive noise. The latent random variable $\mathbf{S}$ is typically assumed to be a vector in $\mathbb{R}^n$; though in principle, it could be a vector in any space $\mathbb{R}^m$ where $m \leq n$. The individual components of $\mathbf{S}$ are assumed to be independent random variables. A is taken to be a full rank matrix, $A \in \mathbb{R}^{n \times m}$. It will be assumed for simplicity that $m = n$, thus making A invertible. We will further assume that each random variable $\mathbf{S}_i$ has variance 1. Note that this last assumption serves to remove an ambiguity of the problem, since the columns of A could otherwise be chosen to have any scale. As a result of these assumptions, the covariance $\text{Cov}(\mathbf{S})$ becomes the identity matrix I.

As discussed in the introduction, most ICA algorithms can be broken down into 2 steps. In the first step, the independent components are made orthogonal and rescaled such that $\mathbf{X} = R\mathbf{S}$ where R is an orthogonal matrix. This method of decorrelating the independent components is termed whitening. In the second step, the columns of R (which correspond to independent components) up to sign and order are found.

In the noisy case the main challenge is presented by Step 1, as Step 2 for kurtosis-based methods is naturally invariant to Gaussian noise. Since additive Gaussian noise affects the covariance matrix

$\mathrm{Cov}(A\mathbf{S}+\boldsymbol{\eta})$, PCA based whitening fails to orthogonalize the independent components. It was observed in [1] that a variation on step 1 could be used. It is enough to make the independent components orthogonal without giving them the same scale. Whereas true whitening sets $\mathbf{X} = R\mathbf{S}$, we replace R with RD such that R is orthogonal and D is a diagonal scaling matrix. Thus, following [1], *quasi-whitening*[2] can be defined as follows:

Definition 3.1. *A quasi-whitening matrix is a matrix W such that $WA = RD$ for some orthogonal matrix R and nonsingular diagonal matrix D.*

We shall now state our main result. Let $\mathbf{e}_1, \cdots, \mathbf{e}_n$ be the canonical vectors that form a basis for the space spanned by the random vector $\mathbf{S}$. Let $\kappa_{\min} = \min_i(|\kappa_4(\mathbf{S}_i)|)$, $\kappa_{\max} = \max_i(|\kappa_4(\mathbf{S}_i)|)$, and $\mu_k = \max_i(\mathbb{E}[\mathbf{S}_i^k])$. Let A_i denote the i^{th} column of matrix A. For clarity of the presentation, we use the following machine model for the running time: a random access machine that allows the following exact arithmetic operations over real numbers in constant time: addition, substraction, multiplication, division and square root.

Theorem 3.2. *Let $\epsilon > 0$ and $\delta \in (0,1)$. Given*

$$N = O\left(\frac{n^{10}\kappa(A)^{16}}{\epsilon^2\delta}\frac{\kappa_{\max}^2}{\kappa_{\min}^4}\left(\mu_8 + \frac{\sigma_{\boldsymbol{\eta}}^8}{\sigma_{\min}(A)^8}\right)\right)$$

samples of $\mathbf{X} = A\mathbf{S} + \boldsymbol{\eta}$ we can compute, in time polynomial in N and n, an approximate quasi-whitening matrix $\hat{B}$ so that with probability at least $1-\delta$ over the sample we have

1. *For $i \neq j$,*

$$-\epsilon \leq \frac{\langle \hat{B}^{-1}A\mathbf{e}_i, \hat{B}^{-1}A\mathbf{e}_j\rangle}{\|\hat{B}^{-1}A\mathbf{e}_i\|_2\|\hat{B}^{-1}A\mathbf{e}_j\|_2} \leq \epsilon \tag{1}$$

2. *The length of $\mathbf{e}_j$ is scaled under the transformation $\hat{B}^{-1}A$ as:*

$$(1-\epsilon)\|A_i\|_2^2 \leq \|\hat{B}^{-1}A\mathbf{e}_j\|_2^2 \leq (1+\epsilon)\|A_i\|_2^2 \tag{2}$$

In simpler words, quasi-whitening approximately orthogonalizes the independent components of $\mathbf{X}$ and scales the independent components based on the lengths of the columns of A.

We note that existing cumulant-based methods already employed for step 2 in ICA can be modified in reasonably straightforward ways to work under quasi-whitening. Several popular ICA algorithms including JADE [3] and the kurtosis based implementation of FastICA [8, 10] are implemented using cumulants. Since higher order cumulants ignore Gaussian noise, this allows for the creation of a class of new algorithms which are resistant to additive Gaussian noise. In the special case where each $\mathbf{S}$ is drawn uniformly from $\{-1,1\}^n$, this has been done by Arora et al in [1].

To see the validity of fourth cumulant based algorithms for the second step of ICA in the presence of Gaussian noise, we draw from Observation 2 of Frieze et al in [6]. An interpretation of the statement and proof is that given $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$ such that each $\alpha_i \neq 0$ and a function of the form $G(\mathbf{v}) = \sum_{i=1}^n \mathbf{v}_i^4\alpha_i$ such that $\mathbf{v}$ is drawn from the unit sphere, when there exists some $\alpha_i > 0$, a complete list of local maxima of $G(\mathbf{v})$ is given by $\{\pm\mathbf{e}_i : \alpha_i > 0\}$ (where $\mathbf{e}_i$ is the ith canonical vector). Similarly, when there exists some $\alpha_i < 0$, a complete list of local minima of $G(\mathbf{v})$ is given by $\{\pm\mathbf{e}_i : \alpha_i < 0\}$ Using the properties of cumulants, it follows that given $\mathbf{v} \in \mathbb{R}^n$ drawn from the unit sphere,

$$\kappa_4(\mathbf{v}\cdot\mathbf{S}) = \sum_{i=1}^n \mathbf{v}_i^4\kappa_4(\mathbf{S}_i), \tag{3}$$

[2]Hyvärinen had a different definition of quasi-whitening in [9].

where $\kappa_4(\mathbf{S}_i)$ takes on the role of α_i. As such, any algorithm which maximizes $|\kappa_4(\mathbf{v}\cdot\mathbf{S})|$ or alternatively $\kappa_4(\mathbf{v}\cdot\mathbf{S})^2$ will find the canonical vectors. Of course, one cannot work in the coordinate system of $\mathbf{S}$, but under the assumption of orthogonality provided, one can instead maximize $|\kappa_4(\mathbf{u}\cdot\mathbf{X})|$ where $\mathbf{u}$ is drawn form the unit sphere since

$$\kappa_4(\mathbf{u}\cdot\mathbf{X}) = \kappa_4(\mathbf{u}\cdot(RD\mathbf{S})) = \kappa_4((R^T\mathbf{u})\cdot(D\mathbf{S}))$$

using that additive Gaussian noise is ignored by cumulants. $D\mathbf{S}$ is simply a rescaling of $\mathbf{S}$, and $\kappa_4(\mathbf{S}_i)$ can be replaced by $\kappa_4(d_{ii}\mathbf{S}_i)$ in equation (3). Using the change of variable $\mathbf{v} = R^T\mathbf{u}$, any locally maximal value for $\mathbf{u}$ will correspond to a column of R, thus recovering a component $\mathbf{S}_i$ up to scaling and noise. In [6], Observation 2 summarizes a very similar result in the case of true whitening without additive Gaussian noise using the fourth moment instead of fourth cumulant, and a mostly correct efficient algorithm and analysis is provided for the fourth moment based on this observation.

4 How to Achieve Quasi-Whitening

Recall that $Q_{\mathbf{X}}$ denotes the fourth cumulant tensor of the observed variable $\mathbf{X}$, with $ijkl^{th}$ entry:

$$(Q_{\mathbf{X}})_{ijkl} = \mathrm{Cum}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l),$$

and define an operation of tensors on matrices $\mathbb{T}\times\mathbb{R}^{n\times n}\to\mathbb{R}^{n\times n}$ by:

$$(Q_{\mathbf{X}}\circ M)_{ij} = \sum_{k,l=1}^{n}\mathrm{Cum}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l)m_{lk}.$$

Before proceeding with the argument leading to the construction of a quasi-whitening matrix, it is worth making several observations about this operation. First, the operation can be viewed as matrix-vector multiplication. Use multi-indices α, β such that α runs over (i, j) and β runs over (l, k), and note that by symmetry, $(Q_{\mathbf{X}})_{ijkl} = (Q_{\mathbf{X}})_{ijlk} = (Q_{\mathbf{X}})_{\alpha\beta}$. Under this flattening of the tensor $Q_{\mathbf{X}}$, the operation becomes matrix-vector multiplication with M taking on the role of the vector using $m_{lk} = m_\alpha$.

The following Lemma describes how the cumulant tensor transforms under a linear change of variable:

Lemma 4.1. *Given a random vector-valued variable $\mathbf{Y}\in\mathbb{R}^n$ and matrices $B, M\in\mathbb{R}^{n\times n}$, then $Q_{B\mathbf{Y}}\circ M = B(Q_{\mathbf{Y}}\circ(B^TMB))B^T$.*

Proof. The proof follows primarily from the multilinearity of cumulants:

$$\begin{aligned}
(Q_{B\mathbf{Y}}\circ M)_{ij} &= \sum_{k,l=1}^{n}\mathrm{Cum}((B\mathbf{Y})_i, (B\mathbf{Y})_j, (B\mathbf{Y})_k, (B\mathbf{Y})_l)m_{lk}\\
&= \sum_{k,l=1}^{n}\sum_{q,r,s,t=1}^{n}\mathrm{Cum}(b_{iq}\mathbf{Y}_q, b_{jr}\mathbf{Y}_r, b_{ks}\mathbf{Y}_s, b_{lt}\mathbf{Y}_t)m_{lk}\\
&= \sum_{k,l=1}^{n}\sum_{q,r,s,t=1}^{n}b_{iq}b_{jr}\mathrm{Cum}(\mathbf{Y}_q, \mathbf{Y}_r, \mathbf{Y}_s, \mathbf{Y}_t)b_{lt}m_{lk}b_{ks}\\
&= \sum_{q,r,s,t=1}^{n}b_{iq}b_{jr}\mathrm{Cum}(\mathbf{Y}_q, \mathbf{Y}_r, \mathbf{Y}_s, \mathbf{Y}_t)(B^TMB)_{ts}\\
&= \sum_{q,r=1}^{n}b_{iq}b_{jr}(Q_{\mathbf{Y}}\circ(B^TMB))_{qr},
\end{aligned}$$

which can be equivalently written as $Q_{B\mathbf{Y}}\circ M = B(Q_{\mathbf{Y}}\circ(B^TMB))B^T$. ∎

The above ideas will be useful both in constructing a quasi-whitening matrix in the noiseless case, as well as in finding an estimate to a quasi-whitening matrix from data. What follows is the construction of a quasi-whitening matrix when one knows the cumulant tensor exactly.

Lemma 4.2. *Let M be an arbitrary matrix. Then, $Q_{\mathbf{X}} \circ M = ADA^T$ where D is a diagonal matrix with q^{th} entry $d_{qq} = \kappa_4(\mathbf{S}_q)A_q^T M A_q$.*

Proof. This proof will proceed by simplifying $Q_{\mathbf{X}} \circ M$ using the properties of cumulants.

$$\begin{aligned}
(Q_{\mathbf{X}} \circ M)_{ij} &= \sum_{k,l=1}^{n} \text{Cum}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l) m_{lk} \\
&= \sum_{k,l=1}^{n} \text{Cum}\left(\sum_{q=1}^{n} A_{iq}\mathbf{S}_q + \boldsymbol{\eta}_i, \sum_{q=1}^{n} A_{jq}\mathbf{S}_q + \boldsymbol{\eta}_j, \right. \\
&\qquad \left. \sum_{q=1}^{n} A_{kq}\mathbf{S}_q + \boldsymbol{\eta}_k, \sum_{q=1}^{n} A_{lq}\mathbf{S}_q + \boldsymbol{\eta}_l \right) m_{lk} \\
&= \sum_{k,l=1}^{n} \sum_{q=1}^{n} \text{Cum}(A_{iq}\mathbf{S}_q, A_{jq}\mathbf{S}_q, A_{kq}\mathbf{S}_q, A_{lq}\mathbf{S}_q) m_{lk} \\
&= \sum_{k,l=1}^{n} \sum_{q=1}^{n} A_{iq}A_{jq}A_{kq}A_{lq}\text{Cum}(\mathbf{S}_q, \mathbf{S}_q, \mathbf{S}_q, \mathbf{S}_q) m_{lk},
\end{aligned}$$

where the last two equalities come from the independence, multilinearity, and vanishing Gaussian properties. Switching into univariate cumulant notation and rearranging summations yields:

$$\begin{aligned}
(Q_{\mathbf{X}} \circ M)_{ij} &= \sum_{q=1}^{n} A_{iq}A_{jq}\kappa_4(\mathbf{S}_q) \sum_{k,l}^{n} A_{lq} m_{lk} A_{kq} \\
&= \sum_{q=1}^{n} A_{iq}A_{jq}\kappa_4(\mathbf{S}_q) A_q^T M A_q
\end{aligned}$$

which has matrix form:

$$Q_{\mathbf{X}} \circ M = ADA^T$$

where D is a diagonal matrix with diagonal entries $d_{qq} = \kappa_4(\mathbf{S}_q)A_q^T M A_q$. □

Theorem 4.3. *Let M be the matrix $(Q_{\mathbf{X}} \circ I)^{-1}$. Let B be a factorization matrix such that $BB^T = Q_{\mathbf{X}} \circ M$. Then, B^{-1} is a Quasi-Whitening matrix.*

Proof. Applying Lemma 4.2 gives $Q_{\mathbf{X}} \circ I = AD'A^T$ with $d'_{qq} = \kappa_4(\mathbf{S}_q)A_q \cdot A_q$. Note that $M = (A^T)^{-1}D'^{-1}A^{-1}$. Applying Lemma 4.2 a second time yields $Q_{\mathbf{X}} \circ M = ADA^T$ where $d_{qq} = \kappa_4(\mathbf{S}_q)A_q^T M A_q$ gives the diagonal elements of D. Manipulating d_{qq} yields:

$$\begin{aligned}
d_{qq} &= \kappa_4(\mathbf{S}_q)A_q^T(A^T)^{-1}D'^{-1}A^{-1}A_q \\
&= \kappa_4(\mathbf{S}_q)\mathbf{e}_q^T(D')^{-1}\mathbf{e_q} \\
&= \kappa_4(\mathbf{S}_q)[\kappa_4(\mathbf{S}_q)A_q \cdot A_q]^{-1} \\
&= \frac{1}{\|A_q\|_2^2}
\end{aligned}$$

Note that d_{qq} is a positive number for each diagonal entry of D. $D^{1/2}$ exists and can be uniquely defined by taking the positive square root of all diagonal entries. Letting B be any factorization

matrix such that $BB^T = Q_{\mathbf{X}} \circ ((Q_{\mathbf{X}} \circ I)^{-1}) = ADA^T$, then $I = B^{-1}AD^{1/2}(B^{-1}AD^{1/2})^T$ gives that $B^{-1}AD^{1/2} = R$ for some orthogonal matrix R. Hence, $B^{-1}A = RD^{-1/2}$ gives that B^{-1} is a quasi-whitening matrix. □

5 Estimation of Cumulants

So far we have shown that given exact knowledge of the fourth order cumulant tensor for the random variable $\mathbf{X} = A\mathbf{S} + \boldsymbol{\eta}$, it is possible to find a quasi-whitening matrix B^{-1} such that $B^{-1}A = RD$ for some orthogonal and diagonal matrices R and D respectively. In practice, one does not have exact knowledge of the cumulant tensor, and the cumulant tensor thus needs to be estimated from samples. Cumulants can be estimated using k-statistics, which are unbiased estimates of cumulants. k-statistics have been studied within the statistics community, and are discussed in the chapter 4 of [14]. For the fourth order cumulant tensor, given random variables $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$, the k-statistic $k(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l)$, which estimates $\text{Cum}(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l)$, is:

$$k(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l) = \frac{1}{N} \sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)\mathbf{y}_i^{(r)}\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)},$$

where ϕ is a function invariant under permutations of its indices defined by $\phi(i,i,i,i) = 1$, $\phi(i,i,i,j) = \phi(i,i,j,j) = -1/(N-1)$, $\phi(i,i,j,k) = 2/[(N-1)(N-2)]$, and $\phi(i,j,k,l) = -6/[(N-1)(N-2)(N-3)]$ when $i,j,k,l \in [N]$ are distinct [14].

k-statistics share several important properties with the cumulant tensors that they estimate. The k-statistic is symmetric in that $k(X_i, X_j, X_k, X_l)$ is invariant under reordering of indices, and it is also multilinear. Multilinearity is shown for the fourth k-statistic in the following Lemma.

Lemma 5.1. *The k-statistic transforms multilinearly.*

Proof. There are 2 properties of multilinearity. For simplicity of notation, they will be only shown on the first coordinate of the k-statistic function. Let $\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l, \mathbf{Z}_i$ be random variables, and let $c \in \mathbb{R}$. Then

1. The additivity portion of multilinearity comes from:

$$\begin{aligned}
&k(\mathbf{Y}_i + \mathbf{Z}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l) \\
&= \frac{1}{N} \sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)(\mathbf{y}_i^{(r)} + \mathbf{z}_i^{(r)})\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)} \\
&= \frac{1}{N}\left[\sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)\mathbf{y}_i^{(r)}\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)} + \sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)\mathbf{z}_i^{(r)}\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)}\right] \\
&= k(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l) + k(\mathbf{Z}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l).
\end{aligned}$$

2. The multiplicative portion of multilinearity comes from:

$$\begin{aligned}
k(c\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l) &= \frac{1}{N} \sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)c\mathbf{y}_i^{(r)}\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)} \\
&= c\frac{1}{N} \sum_{r,s,t,u=1}^{N} \phi(r,s,t,u)\mathbf{y}_i^{(r)}\mathbf{y}_j^{(s)}\mathbf{y}_k^{(t)}\mathbf{y}_l^{(u)} \\
&= k(\mathbf{Y}_i, \mathbf{Y}_j, \mathbf{Y}_k, \mathbf{Y}_l).
\end{aligned}$$

□

These multilinearity properties imply that

$$\begin{aligned}
&k(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k, \mathbf{X}_l) \\
&= \sum_{qrst} k(A_{iq}(\mathbf{S} + A^{-1}\boldsymbol{\eta})_q, A_{jr}(\mathbf{S} + A^{-1}\boldsymbol{\eta})_r, A_{ks}(\mathbf{S} + A^{-1}\boldsymbol{\eta})_s, A_{lt}(\mathbf{S} + A^{-1}\boldsymbol{\eta})_t) \\
&= \sum_{qrst} A_{iq}A_{jr}A_{ks}A_{lt}k(\mathbf{S}_q + (A^{-1}\boldsymbol{\eta})_q, \mathbf{S}_r + (A^{-1}\boldsymbol{\eta})_r, \mathbf{S}_s + (A^{-1}\boldsymbol{\eta})_s, \mathbf{S}_t + (A^{-1}\boldsymbol{\eta})_t).
\end{aligned}$$

As such, Lemma 4.1 applies also to k-statistic estimates of random variables. In particular, it is possible to think of the k-statistic tensor associated with the random variable $\mathbf{X}$ as being generated from an unobserved k-statistic tensor from the latent samples of $\mathbf{S} + A^{-1}\boldsymbol{\eta}$. We can work directly with the random variable $\mathbf{S} + A^{-1}\boldsymbol{\eta}$ for the purposes of error analysis. This will be a natural approach since the difficulty of the problem relies partially on the kurtosis of the latent distribution for $\mathbf{S} + A^{-1}\boldsymbol{\eta}$.

Let μ_k represent $\max_i \mathbb{E}[\mathbf{S}_i^k]$. By assumption, $\mu_1 = 0$ and $\mu_2 = 1$. Let $\boldsymbol{\eta}^*$ denote $A^{-1}\boldsymbol{\eta}$. Let

$$\sigma_{\boldsymbol{\eta}^*} = \max_{\|u\|=1} (\sqrt{\mathbf{u}^T \Sigma_{\boldsymbol{\eta}^*} \mathbf{u}})$$

where $\Sigma_{\boldsymbol{\eta}^*}$ is the covariance matrix of $\boldsymbol{\eta}^*$. The error induced by estimating the latent fourth cumulant tensor $k_{\mathbf{S}+\boldsymbol{\eta}^*}$ from a sample can be bounded using the following 2 Lemmas:

Lemma 5.2. *Let $\mathbf{Z} = \mathbf{S} + \boldsymbol{\eta}^*$. Then,*

$$\mathrm{Var}(k(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l)) = O(\frac{\max_{i\in[n]} \mathbb{E}[Z_i^8]}{N}).$$

Proof. In order to save space, it will be useful to use multi-index notation. In particular, taking $I = (i_1, i_2, i_3, i_4) \in [n]^4$ and $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in [N]^4$, $\phi_\alpha \mathbf{z}_I^{(\alpha)}$ will be denote

$$\phi(\alpha_1, \alpha_2, \alpha_3, \alpha_4)\mathbf{z}_{i_1}^{(\alpha_1)}\mathbf{z}_{i_2}^{(\alpha_2)}\mathbf{z}_{i_3}^{(\alpha_3)}\mathbf{z}_{i_4}^{(\alpha_4)}$$

Further, the set $\alpha \cap \beta$ will be defined as:

$$\alpha \cap \beta = \{\alpha_i : \alpha_i = \beta_j \text{ for some pair } (i, j)\}.$$

Keeping these notations in mind, we can proceed with the proof. Let $I \in [n]^4$.

$$\begin{aligned}
&\mathrm{Var}(k(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \mathbf{Z}_{i_3}, \mathbf{Z}_{i_4})) \\
&= \mathbb{E}\left[\left(\frac{1}{N}\sum_{\alpha\in[N]^4} \phi_\alpha \mathbf{z}_I^{(\alpha)}\right)^2\right] - \mathbb{E}\left[\frac{1}{N}\sum_{\alpha\in[N]^4} \phi_\alpha \mathbf{z}_I^{(\alpha)}\right]^2 \\
&= \frac{1}{N^2}\sum_{\alpha\in[N]^4}\sum_{\beta\in[N]^4} \mathbb{E}[\phi_\alpha \mathbf{z}_I^{(\alpha)}\phi_\beta \mathbf{z}_I^{(\beta)}] - \frac{1}{N^2}\sum_{\alpha\in[N]^4}\sum_{\beta\in[N]^4} \mathbb{E}[\phi_\alpha \mathbf{z}_I^{(\alpha)}]\mathbb{E}[\phi_\beta \mathbf{z}_I^{(\beta)}] \\
&= \frac{1}{N^2}\sum_{\alpha\in[N]^4}\sum_{\substack{\beta\in[N]^4 \\ \alpha\cap\beta\neq\emptyset}} \mathbb{E}[\phi_\alpha \mathbf{z}_I^{(\alpha)}\phi_\beta \mathbf{z}_I^{(\beta)}] - \frac{1}{N^2}\sum_{\alpha\in[N]^4}\sum_{\substack{\beta\in[N]^4 \\ \alpha\cap\beta\neq\emptyset}} \mathbb{E}[\phi_\alpha \mathbf{z}_I^{(\alpha)}]\mathbb{E}[\phi_\beta \mathbf{z}_I^{(\beta)}] \\
&\leq \frac{1}{N^2}\sum_{\alpha\in[N]^4} \phi_\alpha \sum_{\substack{\beta\in[N]^4 \\ \alpha\cap\beta\neq\emptyset}} \phi_\beta \mathbb{E}[\mathbf{z}_I^{(\alpha)}\mathbf{z}_I^{(\beta)}]. \qquad (4)
\end{aligned}$$

Equation (4) contains the essence of the argument. However, in order to complete the argument, several facts need to be demonstrated. First, it needs to be seen that $\left|\mathbb{E}[\mathbf{z}_I^{(\alpha)}\mathbf{z}_I^{(\beta)}]\right| \leq \max_i(\mathbb{E}[\mathbf{Z}_i^8])$. To see this, use the Cauchy-Schwartz inequality on random variables Y_1, Y_2 to get:

$$\mathbb{E}[Y_1 Y_2] \leq \max(\mathbb{E}[Y_1^2], \mathbb{E}[Y_2^2]) \tag{5}$$

Applying this fact recursively yields that $\left|\mathbb{E}[\mathbf{z}_I^{(\alpha)}\mathbf{z}_I^{(\beta)}]\right| \leq \max_i(\mathbb{E}[\mathbf{Z}_i^8])$.

The second difficulty that arises is seeing how limiting oneself to samples in which $\alpha \cap \beta \neq \emptyset$ restricts the summation. First, let $\text{dist}(\beta)$ denote the number of distinct indices in β. If $c = \text{dist}(\beta)$, then there are $\binom{N}{c}$ choices of index values that can be used to generate β, of which $\binom{N-4}{c}$ certainly do not intersect α. As such,

$$\frac{\binom{N}{c} - \binom{N-4}{c}}{\binom{N}{c}}$$

gives an upper bound on the fraction of index sets in which $\beta \cap \alpha \neq \emptyset$ when $\text{dist}(\beta) = c$. Finally, noting that $|\phi_\beta| \leq 7/(N^{\text{dist}(\beta)-1})$ for sufficiently large N and that $\sum_{\alpha \in [N]^4} \phi_\alpha = O(N)$, we have sufficient tools with which to proceed from (4):

$$\begin{aligned}
\text{Var}(k(\mathbf{Z}_{i_1}, \mathbf{Z}_{i_2}, \mathbf{Z}_{i_3}, \mathbf{Z}_{i_4})) &\leq \frac{1}{N^2}\left|\sum_{\alpha\in[N]^4} \phi_\alpha \sum_{c=1}^{4} \sum_{\substack{\text{dist}(\beta)=c \\ \alpha\cap\beta\neq\emptyset}} \phi_\beta \mathbb{E}[\mathbf{z}_I^{(\alpha)}\mathbf{z}_I^{(\beta)}]\right| \\
&\leq \frac{1}{N^2}\max_{i\in[n]}\mathbb{E}[\mathbf{Z}_i^8] \sum_{\alpha\in[N]^4}|\phi_\alpha| \sum_{c=1}^{4} \sum_{\substack{\text{dist}(\beta)=c \\ \alpha\cap\beta\neq\emptyset}} |\phi_\beta| \\
&\leq \frac{1}{N^2}\max_{i\in[n]}\mathbb{E}[\mathbf{Z}_i^8] \sum_{\alpha\in[N]^4}|\phi_\alpha| \sum_{c=1}^{4} \frac{\binom{N}{c} - \binom{N-4}{c}}{\binom{N}{c}} 7N^{-c+1} \sum_{\text{dist}(\beta)=c} 1 \\
&= O\left(\frac{1}{N^2}\max_{i\in[n]}(\mathbb{E}[\mathbf{Z}_i^8])N^{-1}N^{-c+1}N^c \sum_{\alpha\in[N]^4}|\phi_\alpha|\right) \\
&= O\left(\frac{\max_{i\in[n]}(\mathbb{E}[\mathbf{Z}_i^8])}{N}\right).
\end{aligned}$$

□

Lemma 5.3. *Given $\epsilon, \delta > 0$, the error of each term in the k-statistic tensor for $\mathbf{S} + \mathbf{A}^{-1}\boldsymbol{\eta}$ can be bounded beneath ϵ with confidence $1 - \delta$ using*

$$N = O\left(\frac{n^4}{\epsilon^2\delta}\left(\mu_8 + \frac{\sigma_{\boldsymbol{\eta}}^8}{\sigma_{\min}(A)^8}\right)\right)$$

samples.

Proof. Define $\mathbf{Z} = \mathbf{S} + \boldsymbol{\eta}^*$. Then using Lemma 5.2, $\text{Var}(k(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l)) = O(\frac{1}{N}\max_{q\in[n]}\mathbb{E}[\mathbf{Z}_q^8])$. Using the binomial expansion,

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_q^8] &= \sum_{m=0}^{8}\binom{8}{m}\mathbb{E}[\mathbf{S}_q^m(\boldsymbol{\eta}_q^*)^{8-m}] \\
&= \sum_{m=0}^{4}\binom{8}{2m}\mathbb{E}[\mathbf{S}_q^{2m}(\boldsymbol{\eta}_q^*)^{8-2m}],
\end{aligned}$$

since odd 0-mean Gaussian moments are 0. Using equation (5), we see that the dominant terms are μ_8 and $(\sigma_\eta/\sigma_{\min}(A))^8$. In particular, the cross terms come from $\mathbb{E}[\mathbf{S}_q^{2m}\boldsymbol{\eta}_{8-2m}^*]$. When $m=2$, from (5), it follows that

$$\mathbb{E}[\mathbf{S}_q^4(\boldsymbol{\eta}_q^*)^4] \leq \max(\mu_8, \mathbb{E}[(\boldsymbol{\eta}_q^*)^8]).$$

When $m=1$, then

$$\begin{aligned}\mathbb{E}[\mathbf{S}_q^2(\boldsymbol{\eta}_q^*)^6] &= \mathbb{E}[(\mathbf{S}_q^2(\boldsymbol{\eta}_q^*)^2)(\boldsymbol{\eta}_q^*)^4] \\ &\leq \max(\mathbb{E}[\mathbf{S}_q^4(\boldsymbol{\eta}_q^*)^4)], \mathbb{E}[(\boldsymbol{\eta}_q^*)^8)],\end{aligned}$$

for which $\mathbb{E}[\mathbf{S}_q^4(\boldsymbol{\eta}_q^*)^4)] \leq \max(\mu_8, \mathbb{E}[(\boldsymbol{\eta}_q^*)^8])$ has just been shown. The case $m=3$ can be argued similarly to $m=1$ interchanging the roles of $\mathbf{S}$ and $\boldsymbol{\eta}^*$. Thus, one gets:

$$\mathbb{E}[\mathbf{Z}_q^8] = O(\mu_8 + \mathbb{E}[(\boldsymbol{\eta}_q^*)^8]).$$

For even Gaussian moments, the following equation holds (see for instance [12] section 3.4):

$$\mathbb{E}[\sigma_{\boldsymbol{\eta}^*}^{2k}] = \frac{(2k)!}{k!2^k}\sigma_{\boldsymbol{\eta}^*}^{2k}.$$

It follows that

$$\begin{aligned}\mathbb{E}[\mathbf{Z}_q^8] &= O(\mu_8 + \sigma_{\boldsymbol{\eta}^*}^8) \\ &= O(\mu_8 + \sigma_{\boldsymbol{\eta}}^8/\sigma_{\min}(A)^8).\end{aligned}$$

Chebyshev's inequality states that for a random variable Y, $\Pr(|Y-\mu_Y| \geq c\sigma_Y) \leq \frac{1}{c^2}$. Taking Y to be $k(\mathbf{S}_i+\boldsymbol{\eta}_i^*, \mathbf{S}_j+\boldsymbol{\eta}_j^*, \mathbf{S}_k+\boldsymbol{\eta}_j^*, \mathbf{S}_l+\boldsymbol{\eta}_l^*)$, then since the k-statistic is unbiased, it follows that its expectation is $\mathrm{Cum}(\mathbf{S}_i+\boldsymbol{\eta}_i^*, \mathbf{S}_j+\boldsymbol{\eta}_j^*, \mathbf{S}_k+\boldsymbol{\eta}_j^*, \mathbf{S}_l+\boldsymbol{\eta}_l^*) = \mathrm{Cum}(\mathbf{S}_i, \mathbf{S}_j, \mathbf{S}_k, \mathbf{S}_l)$. c can be chosen such that $\delta/n^4 \geq 1/c^2$. Then, in order to bound the error beneath ϵ, it suffices to satisfy:

$$\epsilon \geq c\sqrt{\mathrm{Var}(k(\mathbf{S}_i+\boldsymbol{\eta}_i^*, \mathbf{S}_j+\boldsymbol{\eta}_j^*, \mathbf{S}_k+\boldsymbol{\eta}_j^*, \mathbf{S}_l+\boldsymbol{\eta}_l^*))},$$

which can be guaranteed by choosing N such that $\epsilon \geq cO((\frac{1}{N}\max_{q\in[n]}(\mathbb{E}[\mathbf{Z}_i^8]))^{1/2})$. This leads to the expression:

$$\begin{aligned} cO\left(\sqrt{\frac{1}{N}\max_{q\in[n]}(\mathbb{E}[\mathbf{Z}_i^8])}\right) &\leq \epsilon \\ O\left(\sqrt{\frac{\mu_8 + (\sigma_{\boldsymbol{\eta}}^8/\sigma_{\min}(A)^8)}{N}}\right)\sqrt{\frac{n^4}{\delta}} &\leq \epsilon \\ N \geq O\left(n^4\frac{\mu_8 + (\sigma_{\boldsymbol{\eta}}^8/\sigma_{\min}(A)^8)}{\epsilon^2\delta}\right).\end{aligned}$$

Applying the union bound, this number of samples is sufficient to guarantee with confidence $1-\delta$ that all terms in the k-statistic tensor for $\mathbf{S}+A^{-1}\boldsymbol{\eta}$ can be bounded beneath ϵ with confidence $1-\delta$. □

6 Error Propagation for Quasi-Whitening

What follows is an analysis for how error propagates throughout the quasi-whitening algorithm. It will be demonstrated that the canonical vectors which act as a basis for the independent components

of $\mathbf{S}$ will remain approximately orthogonal after quasi-whitening given sufficiently many samples. It will be demonstrated that the required number of samples is polynomial in terms of $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, n, $\kappa(A)$, $\frac{\kappa_{\max}}{\kappa_{\min}}$, $\frac{1}{\kappa_{\min}}$, $\frac{\sigma_{\eta}}{\sigma_{\min}(A)}$, and μ_8 where ϵ is the allowable cosine error from orthogonality of the basis vectors, and $1-\delta$ is the confidence of success. Since in the previous section, it was demonstrated that given any $\epsilon > 0$, the sample estimate of the cumulant tensor can have error bounded by ϵ in each term, it suffices to demonstrate that at each step of the algorithm, error does not grow too fast. The confidence is unchanged since only one sample is taken. As a notation, hatted variables shall be used to denote approximations of non-hatted variables. It is assumed that the k-statistic tensor $\hat{Q}_{\mathbf{S}}$ estimate of $Q_{\mathbf{S}}$ is defined from samples of the noisy latent variable $\mathbf{S} + \boldsymbol{\eta}^* = A^{-1}\mathbf{X}$, though for simplicity, $\boldsymbol{\eta}^*$ is suppressed from the subscript notation. (See also the discussion after Lemma 5.1.)

Lemma 6.1. *Given a sample of $\mathbf{X}$, let $\hat{Q}_{\mathbf{X}}$ and $\hat{Q}_{\mathbf{S}}$ be the associated k-statistic estimates for $Q_{\mathbf{X}}$ and $Q_{\mathbf{S}}$ respectively, and let $\hat{M}$ be an estimate for the matrix M such that for some $\epsilon_1, \epsilon_2 > 0$, $\|\hat{Q}_{\mathbf{S}} - Q_{\mathbf{S}}\|_{\max} \leq \epsilon_1$ and $\|\hat{M} - M\|_2 \leq \epsilon_2$. There exists a matrix Y such that $\hat{Q}_{\mathbf{X}} \circ \hat{M} = AYA^T$ and $Q_{\mathbf{X}} \circ M = ADA^T$ where D is the diagonal matrix defined in Lemma 4.2, and the error in the estimate Y is bounded as:*

$$\begin{aligned}\|Y - D\|_2 &\leq \|Y - D\|_F \\ &\leq n^2\|A\|_2^2\|M\|_F\epsilon_1 + \sqrt{n}\epsilon_2\|A\|_2^2(n^2\epsilon_1 + \kappa_{\max})\end{aligned}$$

Proof. Using Lemma 4.1, one gets $\hat{Q}_{\mathbf{X}} \circ \hat{M} = A(\hat{Q}_{\mathbf{S}} \circ (A^T\hat{M}A))A^T$, which gives that Y is well defined, and $Y = (\hat{Q}_{\mathbf{S}} \circ (A^T\hat{M}A))$. By similar reasoning, $D = Q_{\mathbf{S}} \circ (A^TMA)$. In the following investigation of error propagation, the tensors $\hat{Q}_{\mathbf{S}}$ and $Q_{\mathbf{S}}$ will be treated as matrices as described in section 4, and the 2-norm used on the tensors should be interpreted as if the tensor has been flattened to its $n^2 \times n^2$ matrix form. Then:

$$\begin{aligned}\|Y - D\|_F &= \|\hat{Q}_{\mathbf{S}} \circ (A^T\hat{M}A) - Q_{\mathbf{S}} \circ (A^TMA)\|_F \\ &= \|\hat{Q}_{\mathbf{S}} \circ (A^T\hat{M}A) - Q_{\mathbf{S}} \circ (A^T\hat{M}A) + Q_{\mathbf{S}} \circ (A^T\hat{M}A) - Q_{\mathbf{S}} \circ (A^TMA)\|_F \\ &\leq \|\hat{Q}_{\mathbf{S}} - Q_{\mathbf{S}}\|_2\|(A^T\hat{M}A)\|_F + \|Q_{\mathbf{S}}\|_2\|A^T(\hat{M} - M)A\|_F \\ &\leq n^2\epsilon_1\|A\|_2^2\|\hat{M} - M + M\|_F + \kappa_{\max}\|A\|_2^2\sqrt{n}\epsilon_2 \\ &\leq n^2\|A\|_2^2\epsilon_1(\|\hat{M} - M\|_F + \|M\|_F) + \sqrt{n}\kappa_{\max}\|A\|_2^2\epsilon_2 \\ &\leq n^2\|A\|_2^2\epsilon_1(\sqrt{n}\epsilon_2 + \|M\|_F) + \sqrt{n}\kappa_{\max}\|A\|_2^2\epsilon_2 \\ &= n^2\|A\|_2^2\|M\|_F\epsilon_1 + \sqrt{n}\epsilon_2\|A\|_2^2(n^2\epsilon_1 + \kappa_{\max}).\end{aligned}$$

This is also a bound for $\|Y - D\|_2$ based on the standard inequality $\|Y - D\|_2 \leq \|Y - D\|_F$. ∎

Lemma 6.1 above bounds the error growth from tensor operations while placing all error on the diagonal matrix. The next goal is to demonstrate that taking the inverse of a matrix has reasonable error propagation properties. The following Lemma (a portion of Theorem 2.5 from [18]) will be useful:

Lemma 6.2. *Let $\|\cdot\|$ be any consistent matrix norm. Given a matrix C and a matrix perturbation E such that $\|C^{-1}E\| < 1$, and given $\tilde{C} = C + E$, then*

$$\frac{\|\tilde{C}^{-1} - C^{-1}\|}{\|C^{-1}\|} \leq \frac{\|C^{-1}E\|}{1 - \|C^{-1}E\|}$$

From this Lemma, it follows immediately that if $\|E\|_2 \leq 1/(2\|C^{-1}\|_2)$, then

$$\|\tilde{C}^{-1} - C^{-1}\|_2 \leq 2\|C^{-1}\|_2^2\|E\|_2 \tag{6}$$

The main result of this paper is contained in Theorem 3.2, which we prove now.

Proof of Theorem 3.2. This proof is long, and is split into 3 parts. In the first part, the preceding Lemmas are used to propagate error from the estimated latent tensor $Q_{\mathbf{S}}$. Then, a bound on the number of samples required to bound within ϵ the cosine and scaling errors for the basis for the independent subspace from equations (1) and (2) is stated. Finally, it is demonstrated that the bound on angular error is correct.

Let N be a sample size to be chosen later as a function of an arbitrary parameter $\eta > 0$, so that with probability $1 - \delta$ we have $\|\hat{Q}_{\mathbf{S}} - Q_{\mathbf{S}}\|_{\max} < \eta$. Then, let $D' = \text{diag}(\kappa_4(\mathbf{S}_1)\|A_1\|^2, \cdots, \kappa_4(\mathbf{S}_n)\|A_n\|^2)$ be the same as in the proof of Theorem 4.3. By Lemma 4.2, $AD'A^T = Q_{\mathbf{X}} \circ I$. Let Y' be the estimate of D' generated as $AY'A^T = \hat{Q}_{\mathbf{X}} \circ I$. Then by Lemma 6.1, it follows that $\|Y' - D'\|_2 < n^{5/2}\|A\|_2^2\eta$. In order to apply equation (6), it is useful to get error bounds for $\|D'^{-1}\|_2$. It can be shown that:

$$\frac{1}{\kappa_{\max}\sigma_{\max}(A)^2} \leq \|D'^{-1}\|_2 \leq \frac{1}{\kappa_{\min}\sigma_{\min}(A)^2}. \tag{7}$$

Then, it follows that using equation (6):

$$\begin{aligned}
\|Y'^{-1} - D'^{-1}\|_2 &\leq 2\|D'^{-1}\|_2^2\|Y' - D'\|_2 \\
&\leq \frac{2n^{5/2}\|A\|_2^2\eta}{\kappa_{\min}^2\sigma_{\min}(A)^4} \\
&= \frac{2n^{5/2}\kappa(A)^2\eta}{\kappa_{\min}^2\sigma_{\min}(A)^2}
\end{aligned} \tag{8}$$

with the restriction that η must be chosen such that $\|Y' - D'\| = n^{5/2}\|A\|_2^2\eta \leq 1/(2\|D'^{-1}\|_2)$. This can be ensured by requiring that $\eta \leq \kappa_{\min}/(2n^{5/2}\kappa(A)^2)$.

Now, let Y and D be defined such that $ADA^T = Q_{\mathbf{X}} \circ (AD'A^T)^{-1}$ and $AYA^T = \hat{Q}_{\mathbf{X}} \circ (AY'A^T)^{-1}$. By Lemma 6.1,

$$\begin{aligned}
\|Y - D\|_2 &\leq n^2\|A\|_2^2\|(AD'A^T)^{-1}\|_F\eta \\
&\quad + \sqrt{n}\|(AY'A^T)^{-1} - (AD'A^T)^{-1}\|_2\|A\|_2^2(n^2\eta + \kappa_{\max}) \\
&\leq n^2\kappa(A)^2\|D'^{-1}\|_F\eta \\
&\quad + \sqrt{n}\kappa(A)^2\|Y'^{-1} - D'^{-1}\|_2(n^2\eta + \kappa_{\max}) \\
&\leq \frac{n^{5/2}\kappa(A)^2}{\kappa_{\min}\sigma_{\min}(A)^2}\eta + \frac{2n^3\kappa(A)^4\kappa_{\max}}{\kappa_{\min}^2\sigma_{\min}(A)^2}\eta + \frac{2n^5\kappa(A)^4}{\kappa_{\min}^2\sigma_{\min}(A)^2}\eta^2
\end{aligned}$$

which follows by applying (7) and (8).

Since $\eta \leq \kappa_{\min}/(2n^{5/2}\kappa(A)^2)$,

$$\|Y - D\|_2 = O\left(\frac{n^3\kappa(A)^4\kappa_{\max}}{\sigma_{\min}(A)^2\kappa_{\min}^2}\eta\right).$$

Once again, it will be necessary to bound $\|D\|_2$ in order to apply equation (6). Using $D = \text{diag}(1/\|A_1\|_2^2, \cdots, 1/\|A_n\|_2^2)$ from the proof of Theorem 4.3, it follows that:

$$\sigma_{\min}(A)^2 \leq \|D^{-1}\|_2 \leq \sigma_{\max}(A)^2.$$

Applying equation (6) yields:

$$\begin{aligned}
\|Y^{-1} - D^{-1}\|_2 &\leq 2\|D^{-1}\|_2^2\|Y - D\|_2 \\
&\leq 2\sigma_{\max}(A)^4 O\left(\frac{n^3\kappa(A)^4\kappa_{\max}}{\sigma_{\min}(A)^2\kappa_{\min}^2}\eta\right) \\
\frac{\|Y^{-1} - D^{-1}\|_2}{\sigma_{\min}(A)^2} &\leq O\left(\frac{n^3\kappa(A)^8\kappa_{\max}}{\kappa_{\min}^2}\eta\right)
\end{aligned}$$

with the restriction that $\|Y - D\|_2 \leq 1/(2\|D^{-1}\|_2)$. Noting that $\|Y - D\|_2 = O\left(\frac{n^3\kappa(A)^4\kappa_{\max}}{\sigma_{\min}(A)^2\kappa_{\min}^2}\eta\right)$ and $1/\|D^{-1}\|_2 \geq 1/\sigma_{\max}(A)^2$, it suffices to restrict $\eta \leq O\left(\frac{\kappa_{\min}^2}{n^3\kappa(A)^6\kappa_{\max}}\right)$.

Since η is arbitrary (except for upper bound restrictions), η can be chosen such that $\frac{\|Y^{-1}-D^{-1}\|_2}{\sigma_{\min}(A)^2} < \frac{\epsilon}{2}$. This can be accomplished taking $\eta = O\left(\frac{\kappa_{\min}^2}{n^3\kappa(A)^8\kappa_{\max}}\epsilon\right)$. This choice is valid, as both restrictions on η are met when $\epsilon \leq 1$. By Lemma 5.3, taking

$$\begin{aligned} N &= O\left(\frac{n^4}{\eta^2\delta}(\mu_8 + (\sigma_{\boldsymbol{\eta}}^8/\sigma_{\min}(A)^8))\right) \\ &= O\left(n^{10}\frac{(\kappa(A)^{16}\kappa_{\max}^2)}{\epsilon^2\delta\kappa_{\min}^4}(\mu_8 + (\sigma_{\boldsymbol{\eta}}^8/\sigma_{\min}(A)^8))\right) \end{aligned}$$

samples suffice to obtain the desired error bound ϵ with probability $1 - \delta$.

The basis in which $\mathbf{S}$ has independent coordinates is the canonical basis. Therefore, the ultimate goal is to show that, with our choice of an approximate quasi-whitening matrix $\hat{B}^{-1}$ below, the canonical vectors stay approximately orthogonal after applying $\hat{B}^{-1}A$. To see this, factorize $\hat{B}\hat{B}^T = \hat{Q}_{\mathbf{X}} \circ (\hat{Q}_{\mathbf{X}} \circ I)^{-1}$. $\hat{B}^{-1}$ is the approximate quasi-whitening matrix, and $\hat{B}\hat{B}^T = AYA^T$ gives that $\hat{B}^{-1}AY^{1/2} = R$ for some orthogonal matrix R, and $\hat{B}^{-1}A = RY^{-1/2}$. Since Y is symmetric, $Y^{-1/2}$ can be taken to be a symmetric matrix. Take $\mathbf{e}_i, \mathbf{e}_j$ to be canonical vectors. Define δ_{ij} to be the delta function such that

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Then with confidence $1 - \delta$,

$$\begin{aligned} \frac{\langle \hat{B}^{-1}A\mathbf{e}_i, \hat{B}^{-1}A\mathbf{e}_j\rangle}{\|A_i\|_2\|A_j\|_2} &= \frac{\mathbf{e}_i^T A^T \hat{B}^{-T}\hat{B}^{-1}A\mathbf{e}_j}{\|A_i\|_2\|A_j\|_2} \\ &= \frac{\mathbf{e}_i^T Y^{-1}\mathbf{e}_j}{\|A_i\|_2\|A_j\|_2} \\ &\in \left(\frac{D_{ij}^{-1}}{\|A_i\|_2\|A_j\|_2} - \frac{\epsilon}{2}, \frac{D_{ij}^{-1}}{\|A_i\|_2\|A_j\|_2} + \frac{\epsilon}{2}\right) \\ &\supset \left(\frac{\delta_{ij}\|A_i\|\|A_j\|}{\|A_i\|\|A_j\|} - \frac{\epsilon}{2}, \frac{\delta_{ij}\|A_i\|\|A_j\|}{\|A_i\|\|A_j\|} + \frac{\epsilon}{2}\right) \\ &= \delta_{ij} \pm \frac{\epsilon}{2}. \end{aligned}$$

Consider the case where $i = j$. Then,

$$\|\hat{B}^{-1}A\mathbf{e}_i\|_2^2 \in \left(1 \pm \frac{\epsilon}{2}\right)\|A_i\|_2^2,$$

which gives equation (2) Consider the case where $i \neq j$. Then,

$$\frac{\langle \hat{B}^{-1}A\mathbf{e}_i, \hat{B}^{-1}A\mathbf{e}_j\rangle}{\|A_i\|_2\|A_j\|_2} \cdot \frac{\|\hat{B}^{-1}A\mathbf{e}_i\|_2\|\hat{B}^{-1}A\mathbf{e}_j\|_2}{\|\hat{B}^{-1}A\mathbf{e}_i\|_2\|\hat{B}^{-1}A\mathbf{e}_j\|_2} \in \pm\frac{\epsilon}{2}$$

$$\begin{aligned} \frac{\langle \hat{B}^{-1}A\mathbf{e}_i, \hat{B}^{-1}A\mathbf{e}_j\rangle}{\|\hat{B}^{-1}A\mathbf{e}_i\|_2\|\hat{B}^{-1}A\mathbf{e}_j\|_2} &\in \pm\frac{\epsilon}{2} \cdot \frac{1}{1 \pm \epsilon/2} \\ &\subset \pm\epsilon \end{aligned}$$

by restricting $\epsilon < \frac{1}{2}$. This gives equation (1), completing the proof. □

7 Acknowledgments

We would like to thank Navin Goyal for useful discussions.

References

[1] S. Arora, R. Ge, A. Moitra, and S. Sachdeva. Provable ICA with unknown Gaussian noise, and implications for Gaussian mixtures and autoencoders, 2012. arXiv:1206.5349.

[2] A. Bell and T. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[3] J. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 362–370. IET, 1993.

[4] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation*. Academic Press, 2010.

[5] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal processing*, 45(1):59–83, 1995.

[6] A. M. Frieze, M. Jerrum, and R. Kannan. Learning linear transformations. In *FOCS*, pages 359–368, 1996.

[7] D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions, 2012.

[8] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[9] A. Hyvarinen. Gaussian moments for noisy independent component analysis. *Signal Processing Letters, IEEE*, 6(6):145–147, 1999.

[10] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[11] T. Jung, S. Makeig, C. Humphries, T. Lee, M. Mckeown, V. Iragui, and T. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

[12] M. Kendall, A. Stuart, and J. K. Ord. *Kendall's advanced theory of statistics. Vol. 1.* Halsted Press, sixth edition, 1994. Distribution theory.

[13] S. Makino, T. Lee, and H. Sawada. *Blind speech separation*. Springer, 2007.

[14] P. McCullagh. *Tensor methods in statistics*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1987.

[15] J. Morton and L.-H. Lim. Principal cumulant component analysis. Preprint.

[16] P. Q. Nguyen and O. Regev. Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures. *J. Cryptology*, 22(2):139–160, 2009.

[17] A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *ICML*, pages 792–799, 2005.

[18] G. W. Stewart and J. G. Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990.

[19] S. S. Vempala and Y. Xiao. Structure from local optima: Learning subspace juntas via higher order pca. *CoRR*, abs/1108.3329, 2011.

[20] A. Yeredor. Blind source separation via the second characteristic function. *Signal Processing*, 80(5):897–902, 2000.